Selección de Centros para Índices en Espacios Métricos

Ariel Lucero, Norma Edith Herrera, Carina Mabel Ruano

Departamento de Informática

Universidad Nacional de San Luis

San Luis, Argentina

{3033402,nherrera,cmruano}@unsl.edu.ar

Contexto

El presente trabajo se desarrolla en el ámbito de la línea Técnicas de Indexación para Datos no Estructurados del Proyecto Tecnologías Avanzadas de Bases de Datos, cuyo objetivo principal es realizar investigación básica en problemas relacionados al manejo y recuperación eficiente de información no tradicional, diseñando nuevos algoritmos de indexación que permitan realizar búsquedas eficientes sobre datos no estructurados.

Resumen

El concepto de búsquedas por similitud, es decir buscar elementos en una base de datos que sean similares o cercanos a uno dado, tiene aplicación en diversas áreas de computación. Las bases de datos que soportan este tipo de consultas pueden ser modelizadas mediante el concepto de espacio métrico. Un espacio métrico es un par $(\mathcal{X}, d)$, donde $\mathcal{X}$ es un conjunto de objetos y d es una función de distancia definida entre ellos que mide cuán diferentes son. El procesamiento de consultas en espacios métricos es un tema de investigación emergente tanto desde el punto de vista de los algoritmos que las implementan como de los índices que las soportan. En este trabajo abordamos el estudio de algoritmos de indexación basados en particiones compactas buscando mejorar la eficiencia de los mismos.

Palabras claves: *Espacios Métricos, Búsquedas por Similitud, Índices, Particiones Compactas.*

1. INTRODUCCIÓN

El concepto de *búsquedas por similitud* o *por proximidad*, es decir buscar elementos de una base de datos que sean similares o cercanos a uno dado, aparece en diversas áreas de computación, tales como reconocimiento de voz, reconocimiento de imágenes, compresión de texto, biología computacional, inteligencia artificial, minería de datos, entre otras.

En [9] se muestra que este problema se puede expresar como sigue: dado un conjunto de objetos $\mathcal{X}$ y una función de distancia d definida entre ellos que mide cuán diferentes son, el objetivo es recuperar todos aquellos elementos que sean similares a uno dado. Esta función d cumple con las propiedades características de una función de distancia: *positividad* ($d(x, y) \geq 0$), *simetría* ($d(x, y) = d(y, x)$) y *desigualdad triangular* ($d(x, y) \leq d(x, z) + d(z, y)$).

El par $(\mathcal{X}, d)$ se denomina ***espacio métrico***. La base de datos será un subconjunto finito $\mathcal{U} \subseteq \mathcal{X}$. En este nuevo modelo de bases de datos, una de las consultas típicas que implica recuperar objetos similares es la *búsqueda por rango*, que denotaremos con $(q, r)_d$. Dado un elemento $q \in \mathcal{X}$, al que llamaremos *query* y un radio de tolerancia r, una búsqueda por rango consiste en recuperar los objetos de la base de datos cuya distancia a q no sea mayor que r, es decir, $(q, r)_d = \{u \in \mathcal{U} : d(q, u) \leq r\}$.

El tiempo total de resolución de una búsque-

da contiene tres términos, a saber: *T= #evaluaciones de $d\times$ complejidad(d) + tiempo extra de CPU + tiempo de I/O* . En muchas aplicaciones la evaluación de la función d es tan costosa que las demás componentes de la fórmula anterior pueden ser despreciadas. Éste es el modelo usado en este trabajo; por consiguiente, nuestra medida de complejidad será la cantidad de evaluaciones de la función de distancia d.

Una forma trivial de resolver una búsqueda por rango es examinando exhaustivamente la base de datos. Para evitar esta situación, se preprocesa la base de datos por medio de un *algoritmo de indexación* con el objetivo de construir una *estructura de datos o índice*, diseñada para ahorrar cálculos en el momento de resolver una búsqueda.

En [9] se presenta un desarrollo unificador de las soluciones existentes en la temática. En dicho trabajo, se muestra que todos los enfoques para la construcción de índices en espacios métricos consisten en particionar el espacio en clases de equivalencia e indexar las clases de equivalencia. Luego, durante la búsqueda, por medio del índice descartar algunas clases, y buscar exhaustivamente en las restantes. La diferencia entre los distintos algoritmos radica en cómo construyen esta relación de equivalencia. Básicamente se pueden distinguir dos grupos: *algoritmos basados en pivotes* y *algoritmos basados en particiones compactas*.

Algoritmos basados en pivotes: en los algoritmos basados en pivotes [1, 3, 5, 7, 8, 9], la relación de equivalencia se define tomando en cuenta la distancia de los elementos de la base a un conjunto preseleccionado de elementos denominados *pivotes*; en este sentido, dos elementos son considerados equivalentes si están exactamente a la misma de distancia de todos los pivotes. El proceso de indexación consiste en seleccionar k pivotes $\{p_1, p_2, \ldots, p_k\}$, y asignar a cada elemento a el vector o firma $\delta(a) = (d(a, p_1), d(a, p_2), \ldots, d(a, p_k))$. Ante una búsqueda $(q, r)_d$, se usa la desigualdad triangular junto con los pivotes para filtrar elementos de la base de datos sin medir su distancia a la query q. Para ello se computa la distancia de q a cada uno de los pivotes p_i, y luego se descartan todos aquellos elementos a, tales que para algún pivote p_i se cumple que $\mid d(q, p_i) - d(a, p_i) \mid > r$. Los elementos no descartados pasan a formar parte de un conjunto de elementos que se comparan directamente con q para determinar si forman o no parte de la respuesta.

Algoritmos basados en particiones compactas: En el caso de los algoritmos basados en particiones compactas [4, 13, 11, 12], la relación de equivalencia se define teniendo en cuenta la cercanía de los elementos a un conjunto preseleccionado de elementos denominados *centros*; en este caso dos elementos son equivalentes si tienen al mismo centro c como su centro más cercano. El objetivo final es dividir el espacio en zonas tan compactas como sea posible. Para ello seleccionan un conjunto de *centros* $\{c_1, c_2, \ldots, c_k\}$ y dividen el espacio asociando a cada centro. La partición asociada a un centro c_i está formada por el conjunto de puntos que tienen a c_i como su centro más cercano. Existen muchos criterios posibles para descartar zonas o particiones durante una búsqueda. Los dos más populares son:

a. Criterio del hiperplano: es el más básico y el que mejor expresa la idea de partición compacta. Básicamente, si c es el centro de la clase $[q]$ (es decir, el centro más cercano a q) entonces la bola con centro q no intersecta $[c_i]$ si $d(q, c) + r < d(q, c_i) - r$. Es decir, si la bola asociada a q no intersecta el hiperplano que divide su centro más cercano c y el centro c_i, entonces cae fuera de la clase de c_i.

b. Criterio del radio de cobertura: en este caso se trata de limitar la clase $[c_i]$ considerando la bola centrada en c_i que contiene todos los elementos de $\mathcal{U}$ que caen en la clase. Definimos el radio de cobertura de c en el espacio $\mathcal{U}$ como $cr(c) = max_{\ u \in [c] \cap \mathcal{U}}\ d(c, u)$. Luego, podemos descartar $[c_i]$ si $d(q, c_i) - r > cr(c_i)$.

Uno de los principales obstáculos en el diseño de buenas técnicas de indexación es lo que se conoce con el nombre de *maldición de la dimensionalidad*. El concepto de dimensionalidad

está relacionado a la dificultad o facilidad de buscar en un determinado espacio métrico. La dimensión intrínseca de un espacio métrico se define en [9] como $\rho = \frac{\mu^2}{2\sigma^2}$, siendo μ y σ^2 la media y la varianza respectivamente de su histograma de distancias. Es decir que, a medida que la dimensionalidad intrínseca crece, la media crece y su varianza se reduce. Esto significa que el histograma de distancia se concentra más alrededor de su media, lo que influye negativamente en los algoritmos de indexación.

La figura 1 da una idea intuitiva de por qué el problema de búsqueda se torna más difícil cuando el histograma es más concentrado. Los histogramas de la figura representan posibles distribuciones de distancias respecto de algún elemento c (histogramas locales respecto de c). Considerando una búsqueda $(q, r)_d$, las áreas sombreadas de la figura muestran los puntos que no podrán descartarse si se utiliza c como centro. Puede observarse que a medida que el histograma se concentra más alrededor de su media, disminuye la cantidad de puntos que pueden descartarse usando como dato $d(c, q)$. Este fenómeno es independiente de la naturaleza del espacio métrico, y nos brinda una forma de cuantificar cuán dura es una búsqueda sobre el mismo.

El procesamiento de consultas en espacios métricos es un tema de investigación emergente tanto desde el punto de vista de los algoritmos que las implementan como de los índices que las soportan. Por esta razón, en este trabajo abordamos el estudio de algoritmos de indexación basados en particiones compactas buscando mejorar la eficiencia de los mismos.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Se sabe que la forma en que se seleccionan los centros afecta en gran medida el desempeño del índice creado. La selección trivial es la random, pero la experiencia marca que aquellas tareas realizadas aleatoriamente pueden mejorarse incorporando alguna política específica. El grupo de centros seleccionados durante la construcción del índice no afecta en absoluto la efectividad del mismo pero es crucial para su eficiencia.

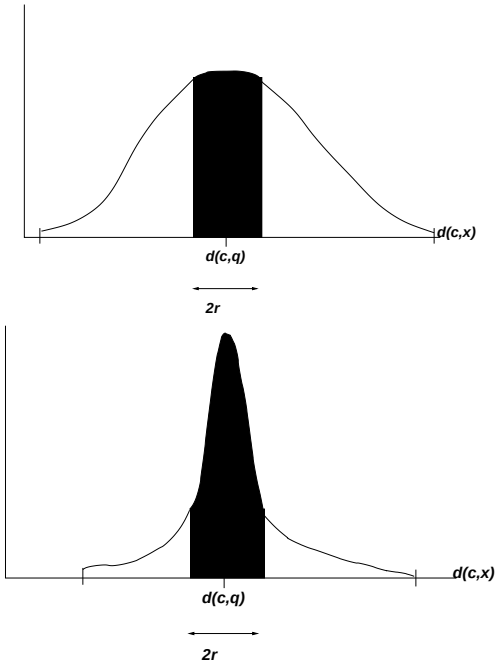

Figura 1: *Histogramas de distancias de baja dimensionalidad (arriba), y de alta dimensionalidad (abajo)*

Una buena selección de centros debería elegir un conjunto de elementos que permitan agilizar las búsquedas mediante el uso de la información obtenida al calcular las distancias entre los centros y la query q. Es decir, se espera que un buen conjunto de centros forme particiones en el espacio que minimicen la cantidad de evaluaciones de la función de distancia necesarias para responder una búsqueda por similitud.

Una característica de un buen conjunto de centros es que sus elementos no estén muy cercanos unos de otros o concentrados en una pequeña zona del espacio ya que, si esto ocurre, es muy probable que la bola de la query intersecte varias zonas del espacio las que no podrán ser descartadas.

En [4] se propone una técnica que intenta evitar elegir centros que estén muy cercanos unos de otros. Para ello se toma una muestra del espacio y luego se seleccionan los elementos de la muestra que están más alejados unos de otros. Para lograr esto, en cada paso, se elije como centro c_i aquel elemento que esté más alejado simultáneamente de $c_1, c_2, \cdots, c_{i-1}$. Este proceso puede resolverse con una complejidad de $O(nm)$ donde n es el número de elementos en la muestra y m es la cantidad de centros deseados.

El proceso de descarte de las búsquedas por si-

militid en espacios métricos es sensible a la distribución de los elementos en el espacio. Es decir, en las zonas del espacio con mayor concentración de elementos, el proceso de descarte de la búsqueda se hace más dificultoso que en otras zonas de menor concentración de elementos. Como ya vimos, una forma de visualizar la distribución de los elementos de un espacio métrico es utilizando histogramas de distancia. En [2] los autores hacen uso de histogramas de distancias para definir el concepto de *núcleo duro* y *núcleo blando* de un espacio métrico. El núcleo duro está formado por aquellos elementos que se ubican en la zona de mayor concentración, que se corresponde con la zona que se encuentra alrededor de la media del histograma de distancias, si el mismo tiene forma de campana de Gauss; el núcleo blando está formado por el resto de los elementos en el espacio métrico.

En [6] se proponen dos técnicas de selección de centros basadas en los conceptos de núcleo duro y núcleo blando. Una de ellas, denominada *closer element* consiste en seleccionar los centros desde el conjunto de elementos pertenecientes al núcleo blando y la otra, denominada *high density zone*, consiste en elegirlos desde el conjunto de elementos pertenecientes al núclo duro. Estas técnicas no utilizan los núcleos duro y blando del espacio métrico (como se sugiere en [2]) sino que en cada paso se crea el histograma local del centro elegido en el paso anterior y la selección del próximo centro se basa sólo en lo que el centro anterior vería como núcleo duro o núcleo blando (que puede no corresponderse con los núcleos reales del espacio). Los autores muestran que *high density zone* es la más competitiva logrando importantes reducciones en la cantidad de evaluaciones de distancias cuando se la compara con una selección aleatoria de centros.

En este trabajo proponemos el estudio de nuevas técnicas de selección de centros para índices métricos basados en particiones compactas. Hemos diseñado hasta el momento dos nuevas políticas de selección, las que explicamos a continuación:

- Se sabe que el histograma local puede ser muy diferente del histograma global del espacio; pero si los histogramas locales de diferentes puntos de referencia son similares, entonces podemos predecir a través de ellos la distribución de los elementos del espacio métrico. Esta es la observación que usan los autores en [2] para proponer como método de detección del núcleo duro, la intersección de varios histogramas locales. Basándonos en esto, la técnica *high density zone* debería mejorar su aproximación al núcleo duro, y en consecuencia mejorar su desempeño, si en lugar de considerar sólo el histograma del último centro elegido c_i considera la intersección de los histogramas de todos los centros elegidos hasta ese momento $c_1, c_2, \cdots, c_i$.

- La técnica *high density zone* se basa en histogramas con forma de campana de Gauss y como se muestra en [10] no todos los histogramas tienen esta forma ; en algunos casos el histograma pueden tener varios máximos locales que por lo general no se corresponden con la media del mismo. Una adaptación de *high density zone* a otros tipos de histogramas es realizar la intersección de aquellas zonas que se encuentren aledañas a todos los máximos locales.

Actualmente nos encontramos implementando estas técnicas para su posterior evaluación experimental sobre índices basados en particiones compactas.

3. RESULTADOS OBTENIDOS/ESPERADOS

Se espera que las técnicas diseñadas resulten más competitivas que *high density zone* dado que estarán aproximando de manera más real el histograma global del espacio métrico y además tendrán en cuenta las distintas formas que puede tener un histograma. Los algoritmos implementados serán evaluados empíricamente utilizando los espacios de prueba ampliamente usados y aceptados por la comunidad científica del área de estudio, los que encuentran disponibles en el sitio de Similarity Search and Applications (SISAP) `http://www.sisap.org`.

4. FORMACIÓN DE RECURSOS HUMANOS

El presente trabajo se desarrolla en el ámbito de la línea *Ténicas de Indexación para Datos no Estructurados* del proyecto *Tecnologías Avanzadas de Bases de Datos* de la Universidad Nacional de San Luis. El desarrollo de este trabajo es parte de un Trabajo Final de la Licenciatura en Ciencias de la Computación de dicha Universidad.

REFERENCIAS

[1] R. Baeza-Yates. Searching: an algorithmic tour. In A. Kent and J. Williams, editors, *Encyclopedia of Computer Science and Technology*, volume 37, pages 331–359. Marcel Dekker Inc., 1997.

[2] R. Baeza-Yates, B. Bustos, E. Chávez, N. Herrera, and G. Navarro. *Clustering in Metric Spaces and Its Application to Information Retrieval*. Kluwer Academic Publishers, 2003. ISBN 1-4020-7682-7.

[3] R. Baeza-Yates, W. Cunto, U. Manber, and S. Wu. Proximity matching using fixed-queries trees. In *Proc. 5th Combinatorial Pattern Matching (CPM'94)*, LNCS 807, pages 198–212, 1994.

[4] S. Brin. Near neighbor search in large metric spaces. In *Proc. 21st Conference on Very Large Databases (VLDB'95)*, pages 574–584, 1995.

[5] W. Burkhard and R. Keller. Some approaches to best-match file searching. *Comm. of the ACM*, 16(4):230–236, 1973.

[6] N. Herrera C. Mendoza Alric. Center selection techniques for metric indexes. *Journal of Computer Science & Technology,*, 7(1):98–104, 2007.

[7] E. Chávez and K. Figueroa. Faster proximity searching in metric data. In *Proceedings of MICAI 2004. LNCS 2972, Springer*, Cd. de México, México, 2004.

[8] E. Chávez, J. Marroquín, and G. Navarro. Fixed queries array: A fast and economical data structure for proximity searching. *Multimedia Tools and Applications (MTAP)*, 14(2):113–135, 2001.

[9] E. Chávez, G. Navarro, R. Baeza-Yates, and J.L. Marroquín. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, September 2001.

[10] E. Chávez, N. Herrera, C. Ruano, and A. Villegas. Funciones de discretización basadas en histogramas de distancia. In *Actas de la Conferencia Latinoamericana de Informática (CLEI'06)*, Santiago, Chile, 2006.

[11] I. Kalantari and G. McDonald. A data structure and an algorithm for the nearest point problem. *IEEE Transactions on Software Engineering*, 9(5):631–634, 1983.

[12] G. Navarro. Searching in metric spaces by spatial approximation. In *Proc. String Processing and Information Retrieval (SPIRE'99)*, pages 141–148. IEEE CS Press, 1999.

[13] J. Uhlmann. Satisfying general proximity/similarity queries with metric trees. *Information Processing Letters*, 40:175–179, 1991.